PONENCIA

Revisión de distintas implementaciones para preservación digital: hacia una propuesta metodológica para la preservación y la auditoría de confiabilidad de RI

MARISA R. DE GIUSTI

GONZALO L. VILLARREAL

Proyecto de Enlace de Bibliotecas-Servicio de Difusión de la Creación Intelectual (PREBI-SEDICI)
Universidad Nacional de La Plata (UNLP)
Centro de Servicios en Gestión de la Información (CESGI)
Comisión de Investigaciones Científicas (CIC)

Resumen

Este trabajo relata la experiencia inicial de prueba de una estructura apta para la preservación de documentos digitales en un archivo o repositorio. Se reconocen numerosos antecedentes de estructuras similares y, entre ellas, se describen brevemente tres experiencias exitosas dedicadas a conectar un repositorio con herramientas capaces de asegurar la preservación digital de los contenidos siguiendo el modelo OAIS, norma ISO 14721 (2012). Tras la descripción de estos tres modelos considerados más relevantes, se relata un prototipo en prueba en los repositorios gestionados en PREBI-SEDICI (UNLP) con las herramientas DSpace, Archivematica y ArchivesSpace, en el que el repositorio en DSpace está encargado del ingreso y la entrega de los contenidos digitales mientras que la estructura de Archivematica realiza las actividades de preservación digital a través de la implementación de un conjunto de microservicios, que actúan sobre una estructura conceptual asimilable al paquete de información (IP) en sus distintas versiones. La estructura física resultante del paquete de información en sus diferentes versiones (SIP, AIP, DIP) incluye archivos, checksum, logs, documentación de la transferencia y metadatos en una estructura XML. Este trabajo no tiene más

pretensiones que mostrar los antecedentes y el inicio de un trabajo de investigación con el objetivo de generar consultas y reflexiones en el contexto latinoamericano, donde estas temáticas son incipientes.

Abstract

This work introduces the initial experience of an infrastructure for digital documents preservation in archives or repositories. Prior backgrounds of similar infraestructures are recognized in this work, and among them three successful experiences are described. These experiences are all aimed to connect a digital repository with different software tools able to ensure digital preservation of repository contents according to OAIS ISO 14721 standard (2012). After the description of the three models, we describe a prototype under development in the repositories supported by PREBI-SEDICI (UNLP), which uses the software tools DSpace, Archivematica and ArchivesSpace. In this prototype, DSpace handles the ingest and delivery of digital contents, while Archivematica performs all the required digital preservation activities. This is achieved through a set of microservices applied to a conceptual structure similar to the information package (IP) in its differents versions (SIP, AIP, DIP). The resulting structure of the IP includes checksums, original files, logs, transfer documentation and XML metadata. The main purpose of this work is to show the background activities already carried out in institutions around the world, and to start a research project aiming to generate ideas and thoughts in the Latin American context.

Introducción

La preservación digital (PD) tiene como propósito asegurar el acceso a largo plazo de contenidos digitales. Existen muchas maneras de alcanzar esta meta: monitoreo de formatos, control de integridad de archivos, migraciones y emulación de entornos, por mencionar algunas. Los repositorios digitales, concebidos como espacios para alojar y difundir grandes cantidades de objetos digitales (OD), requieren conocer profundamente los distintos aspectos de la PD y adoptar las medidas necesarias para asegurar el acceso a largo plazo de los OD que alojan. Para ello, cada organización debe establecer una estructura interna capaz de realizar las distintas actividades para preservar digitalmente sus objetos, lo que requiere implementar sistemas informáticos, adecuar las infraestructuras tecnológicas y establecer los procesos durante todo el ciclo de vida de los OD, que permitan analizarlos y transformarlos a medida que sea necesario.

No existe hoy en día consenso sobre cómo debe implementarse una estructura adecuada que permita asegurar la PD. Sin embargo, existen buenas prácticas, sistemas informáticos y flujos de trabajo que han probado ser muy apropiados para la gran mayoría de las tareas de preservación, y que también resultan centrales para la evaluación y auditoría de un repositorio institucional (RI).

El objetivo central de este trabajo es analizar un número acotado de estructuras recomendadas para preservar y dar acceso a lo largo del tiempo a los OD almacenados en un repositorio o en estructuras aptas para albergarlos. Las tres propuestas analizadas han sido probadas por otras instituciones de gran prestigio, incluso por exitosos proyectos que han reunido varias instituciones, repositorios y archivos, interesados en la preservación digital. En el decurso de los apartados quedará más clara la razón de la configuración propuesta para probar en los repositorios CIC Digital y SEDICI, gestionados por los grupos de trabajo de PREBI-SEDICI.

Adelantando las conclusiones, basta decir que tras un análisis previo de herramientas, al tener las implementaciones facilidades básicas comunes, como agregado de metadatos especiales para preservación y seguimiento del ciclo de vida del OD, la elección atiende a la implementación más común de RI de la región latinoamericana basada en la herramienta de código abierto DSpace. El entorno propuesto para las pruebas podría variarse, pero aquí sí ha intervenido la forma de trabajo propia de los repositorios en gestión que, además, ofrece una estructura más sencilla.

Antecedentes: estructuras e implementaciones exitosas

1. Estructura utilizada en el proyecto SCAPE

Scalable Preservation Environments (SCAPE) es un proyecto coordinado por el Austrian Institute of Technology, financiado por la Unión Europea. El proyecto comenzó en 2012 y finalizó, de acuerdo a lo previsto, a fines de 2015. Reunió a expertos de instituciones que alojan patrimonio cultural, centros de cálculo, laboratorios, universidades e industrias para estudiar aspectos tecnológicos y organizativos de la preservación digital. Los puntos centrales del proyecto fueron:

- El análisis de formatos de ficheros de repositorios
- La descripción formal de planes y políticas de preservación

- La automatización y vitalización de herramientas y procesos escalables
- El control de calidad de procesos de preservación

La arquitectura de referencia propuesta por SCAPE incluye, en el ciclo de vida de la preservación: 1) una instancia de repositorio, que puede generarse sobre DSpace, Eprints o RODA y en la implementación de la estructura de referencia sobre RODA; 2) un monitor de los diferentes procesos internos y externos, encargado además de notificar los riesgos y las oportunidades, especialmente en relación a los aspectos de gestión de la preservación, implementado en SCOUT; 3) un proceso detallado de planeamiento de la preservación realizado a través de PLATO, que desemboca en un plan de preservación que se ingesta en el repositorio y 4) un sistema de gestión del flujo de trabajo que permite la ejecución y sincronización de tareas complejas, como caracterización, migración y otras. Estas tareas son realizadas por diferentes herramientas, integradas al flujo de trabajo del Sistema de Gestión de Flujos de Trabajo TAVERNA. La implementación de referencia de SCAPE sigue un esquema similar al expuesto en la Figura 1.

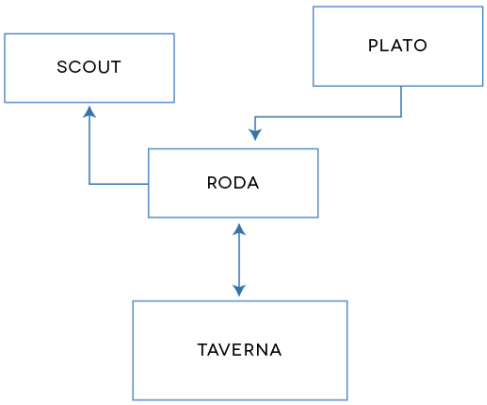

Figura 1. SPE (SCAPE Preservation Environment). Implementación de referencia

Miguel Ferreira y otros (2014) presentan la arquitectura e implementación del proyecto SCAPE y los resultados de su evaluación en relación a la norma ISO 16363, el estándar para auditoría y certificación de repositorios confiables. Más allá de una descripción detallada del modelo, el trabajo señala que efectivamente el esquema/modelo desarrollado es capaz de cumplir con la mayor parte de los requerimientos de dicha norma, haciendo la salvedad de que las métricas vinculadas a la organización que sustenta el repositorio y sus procedimientos exceden lo alcanzable por una herramienta tecnológica. Con

esta visión, el análisis pone fuera de su alcance las métricas de la sección de la norma denominada *Organizational Infrastructure* y mayoritariamente las referidas a la sección *Organizational Structure & Staffing*, excepto lo relativo a las políticas del repositorio, la transparencia y la integridad de los datos, en tanto es capaz de soportar de manera parcial el seguimiento de la gestión intelectual, derechos y restricciones; no obstante, lo parcial está vinculado al problema humano de que no estén los permisos debidamente aclarados o seteados incorrectamente cuando se arma el AIP. El modelo permite y da cuenta completamente de las funciones previstas en las secciones 4.2 y 4.3 de la norma vinculadas a la ingesta, creación del AIP, planeamiento de la preservación, gestión de la preservación y gestión del acceso. Se exceptúan los procesos vinculados a procedimientos que dependen de la institución, por ejemplo aquellos relacionados a las acciones a tomar sobre los AIP, e incluso procedimientos más allá de la documentación vinculados a la tecnología (hardware y software) para cumplir con los requisitos de la comunidad designada o la gestión de riesgos.

El trabajo es sumamente detallado y su lectura muy recomendable para avanzar en los temas de confiabilidad. En el marco de este trabajo, sin embargo, lo más importante está vinculado a estudiar el funcionamiento de la estructura y, naturalmente, la funcionalidad vinculada a las entidades propuestas por la ISO 14721, así como la distribución de esa funcionalidad en el SPE.

Si bien el proyecto ya ha finalizado, todos los materiales generados y la información recopilada a lo largo de cuatro años está compartida libremente en su sitio web.

Acerca de RODA

RODA fue desarrollado para ser un repositorio digital completo, y provee la funcionalidad necesaria para las principales unidades que componen el modelo de referencia OAIS (Faria *et al.*, 2009). RODA implementa todo el flujo de trabajo de ingesta (*ingest*), en el que no sólo valida los SIP sino también se encarga del proceso de negociación entre el archivo y el productor de información. Para el proceso de acceso (*access*) provee diferentes posibilidades de búsqueda y navegación a través de los metadatos, además de visualizaciones y descargas de los OD almacenados. Los componentes de la administración (*Administration*) también fueron desarrollados para permitir a los archivistas modificar los metadatos descriptivos y definir reglas para intervenciones de preservación, como la planificación de controles de

integridad sobre todos los OD almacenados, la iniciación de un proceso de migración o el control de usuarios y/o grupos que están autorizados a ejecutar acciones dentro del repositorio.

El modelo de contenido de RODA es atomístico y muy orientado a PREMIS. Cada entidad intelectual es descrita por un componente EAD (Encoded Archival Description) (Pitti, 1999) de registro de metadatos. Estos registros se organizan jerárquicamente a fin de constituir una descripción de archivo completa, pero manteniéndolos separados dentro del modelo de contenidos de Fedora Commons (Fedora, 2017). Estos componentes EAD son creados a través del mecanismo de enlaces RDF de Fedora, y cada nodo, hoja del árbol jerárquico (Figura 2), es enlazado a un objeto de representación (ejemplo: un objeto Fedora que incluye todos los archivos y *bitstreams* que componen la representación digital). Se mantienen relaciones lógicas entre todos estos objetos, por medio de un conjunto de entidades PREMIS (nodos PO), a fin de conocer la historia y origen (*provenance*) de cada objeto.

Los eventos de preservación que se llevan a cabo se registran como nuevos nodos de eventos de preservación. Algunos eventos especiales, como migraciones de formatos, establecen relaciones adicionales entre dos nodos de representación de preservación (eventos de enlazado). Cada evento de preservación es ejecutado por un agente, que puede ser un usuario del sistema o un evento disparado automáticamente por el software. Como es de esperarse, la información del agente que disparó el evento también se registra dentro del nodo PO.

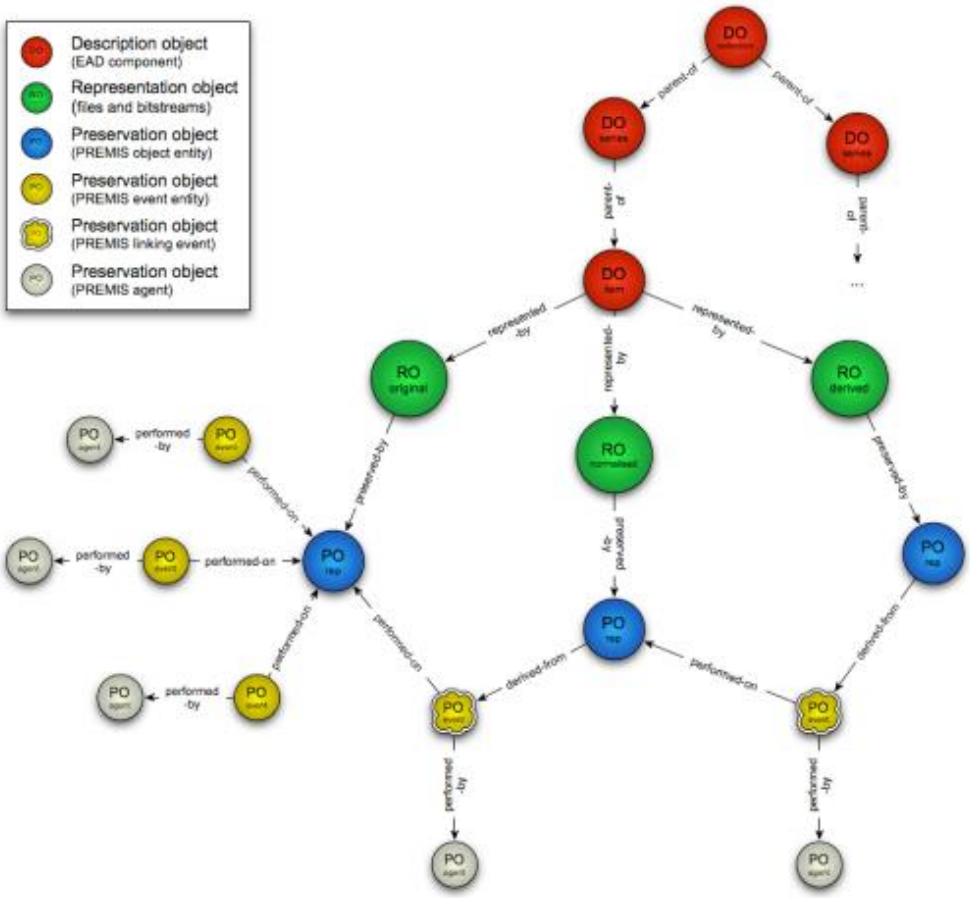

Figura 2. Modelo de contenidos de RODA, con componentes EAD, representaciones digitales y objetos PREMIS

Fuente: Faria et al. (2009)

2. Estructura utilizada por British Columbia University

Lori J. Ashley (2016) considera los riesgos tecnológicos a los que están sometidos los activos digitales, analiza los límites y alcances en el tiempo de un conjunto dado de estrategias y realiza un somero relato del modelo de referencia OAIS, la norma ISO 16363 y algunos tópicos tecnológicos centrales a la preservación: actualización, réplica, migración, emulación, normalización de formatos, etcétera. Esto sirve de introducción teórica al trabajo de Bronwen Sprout y Sarah Romkey (2016), que relata la experiencia de la Biblioteca de la Universidad de British Columbia (UBC), particularmente la implementación del repositorio institucional. Este repositorio fue implementado en DSpace con la colaboración de Artefactual Systems, creadores de Archivematica. Tras un análisis de las prácticas de preservación que se realizaban tanto sobre materiales nacidos digitales como los surgidos a partir de procesos de digitalización, Artefactual observó las

deficiencias, realizó un diagnóstico y una propuesta de estructura a probar en un proyecto piloto, mostrado en la Figura 3.

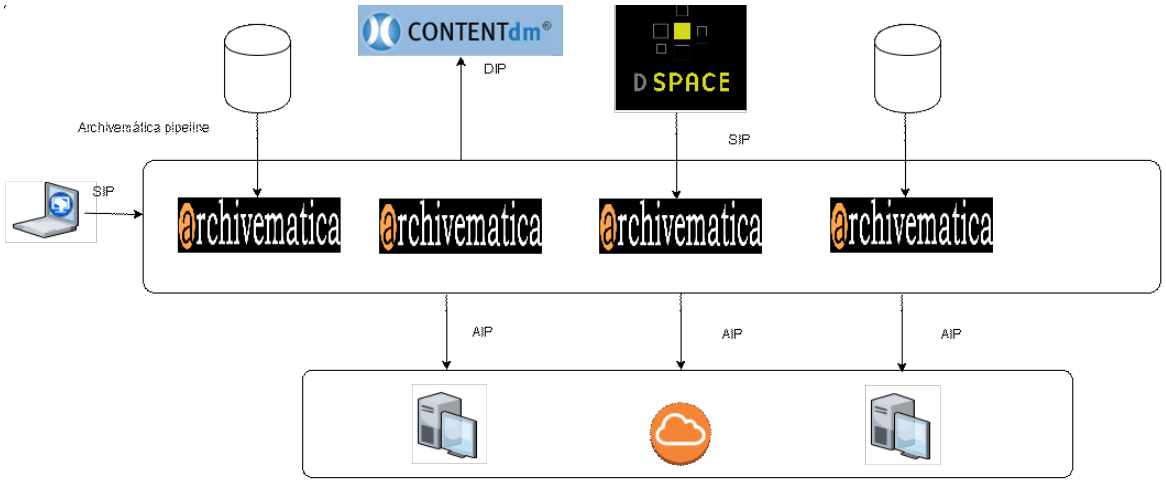

Figura 3. Diagrama general de un *pipeline* de Archivematica

Fuente: Sprout y Romkey (2016)

El esquema de la Figura 3 da cuenta de diversos mecanismos (pudiera haber otros distintos) de ingreso de un SIP en términos de OAIS, es decir de un archivo al *pipeline*, y la ejecución por parte de Archivematica de las funciones descriptas por la estructura abstracta del OAIS que permiten generar un AIP (un paquete preservable) y también un DIP (un paquete entregable) para cualquier configuración; por ejemplo, en la imagen se incluye CONTENTdm, una herramienta enfocada en el almacenamiento y gestión de activos digitales. Cada instancia de Archivematica dentro del *pipeline* puede encargarse del procesamiento de contenido para distintas aplicaciones que lo utilizarán de algún modo. Las distintas instancias del *pipeline* pueden almacenar el paquete preservable en diferentes espacios virtuales, como servidores locales o una nube.

En el trabajo mencionado (Sprout y Romkey, 2016) se relata la experiencia concreta de la implementación de la Universidad de British Columbia que cuenta con un repositorio basado en DSpace, denominado cIRcle. En esta configuración DSpace sirve como herramienta de depósito y acceso (SIP y DIP), pero no se encarga de la generación del paquete preservable AIP, tarea que realiza alguna de las instancias de Archivematica en el pipeline. Es importante observar que esta modalidad no afecta la interfaz del usuario ni su experiencia: el repositorio es el punto de entrada para sus archivos y desde el

repositorio recibe las respuestas a las solicitudes de información que ha ingresado.

Acerca de DSpace

DSpace es un desarrollo de código abierto que permite la implementación de un repositorio, incluyendo la posibilidad de ingreso de archivos de distinto formato, el agregado de metadatos para su catalogación, el almacenamiento, la replicación, la difusión y la entrega ante solicitudes de usuarios del repositorio. Si bien DSpace incluye algunos requisitos del modelo OAIS, no resulta sencillo cumplir con la mayor parte de las funciones que OAIS describe dentro de la entidad denominada *Preservation Planning* (Planeamiento de la Preservación), sobre todo porque esa entidad presupone un comportamiento evolutivo; DSpace sí puede realizar transformaciones, como migraciones de los datos. Las funciones que se pueden llevar adelante desde la administración, asimilables a las de la entidad *Administration*, que es la más compleja del modelo OAIS, están lejos de cumplir con la totalidad de las necesarias para asegurar la preservación (De Giusti y otros, 2012). Particularmente, resulta dificultoso separar agentes y eventos de acuerdo a la descripción del diccionario de datos PREMIS, cuestiones que serían trascendentes a la hora de recuperar o realizar cualquier tarea de transformación sobre los OD. Tampoco los posibles eventos tienen una descripción adecuada, de modo que es difícil seguir el ciclo de vida del OD y asegurar las acciones de preservación necesarias para garantizar su acceso y legibilidad a lo largo del tiempo.

Es esta dificultad, que presenta el software a la hora de la preservación digital (De Giusti, 2016), una de las razones que impulsan a pensar en estructuras como las revisadas en este trabajo.

Acerca de Archivematica

Archivematica es una aplicación de código abierto, basada en estándares reconocidos que asegura el acceso a largo plazo de los archivos digitales. Desarrollada por Artefactual Systems, consta de un conjunto de aplicaciones integradas y herramientas open-source que permiten a los usuarios procesar los OD desde su ingreso (*ingest*) y su almacenamiento hasta la entrega (*access*) siguiendo el modelo ISO-OAIS. Las funcionalidades de Archivematica serán descritas con mayor detalle en el apartado 4.3.

3. Estructura utilizada por Bentley Historical Library (Michigan University)

Tras una revisión de herramientas open-source para software de gestión de activos digitales, la biblioteca de la Universidad de Michigan decidió integrar las funcionalidades de ArchivesSpace, Archivematica y DSpace para lograr un flujo de trabajo a través de una estructura como la que se esquematiza en la Figura 4.

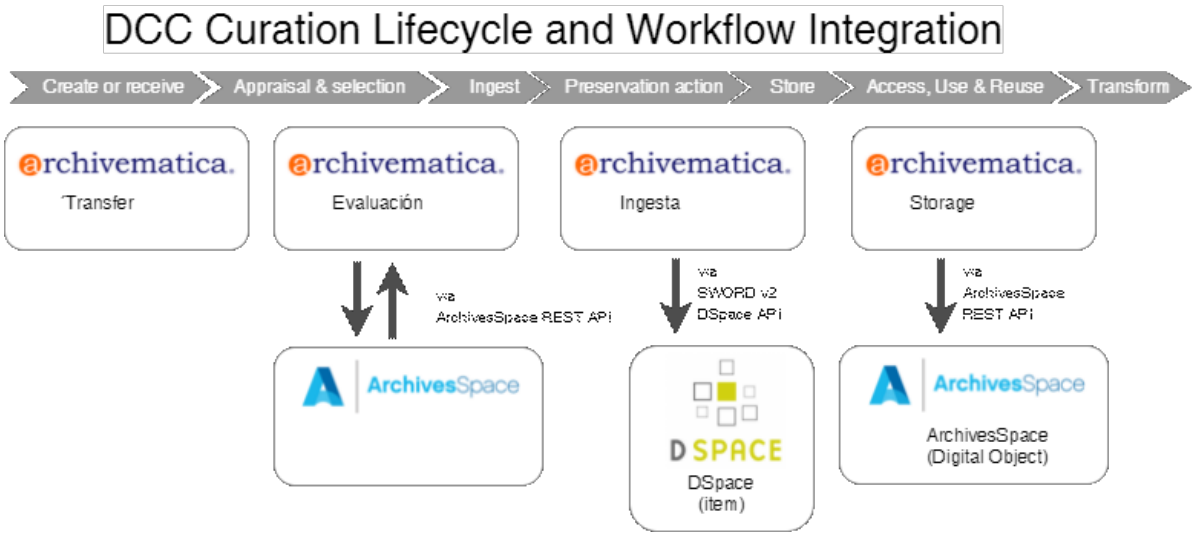

Figura 4. Estructura utilizada por la Biblioteca Histórica de Bentley, Universidad de Michigan

Fuente: Eckard, Pillen & Schallcross (2017)

La estructura propuesta cumple con las necesidades de preservación y acceso a largo plazo para los objetos nacidos digitales según los criterios de la institución. Dentro de la estructura implementada se realizan funciones distribuidas de acuerdo a las necesidades de la institución:

- Facilitar la creación/reutilización de metadatos descriptivos y administrativos en los sistemas de conservación y gestión.
- Simplificar la ingesta y el depósito de contenido en un repositorio de preservación.

Encontrar soluciones para Bentley, pero extrapolables a otras instituciones. Compartir todo el código y la documentación con los archivos y las comunidades de preservación digital.

Acerca de ArchivesSpace

ArchivesSpace es un software abierto de gestión de archivos que permite a las instituciones dar seguimiento a las sesiones de acceso, la gestión de

colecciones así como generar una descripción EAD (Encoded Archival Description). A nivel básico, un documento es un "instrumento de descripción" codificado utilizando EAD, que consta de tres segmentos: uno que proporciona información sobre el instrumento de descripción en sí mismo (título, compilador, fecha de compilación), un segundo componente que incluye lo necesario para la publicación formal del instrumento de descripción, y un tercero que proporciona la descripción del material archivístico, además de la información contextual y administrativa asociada. ArchivesSpace ayuda a realizar las funciones básicas de cualquier archivo: descripción de materiales, gestión de las autoridades de los documentos, gestión de los documentos y de cuestiones relacionadas con ellos (como el número de visualizaciones), e incluso la edición de metadatos dentro de la propia aplicación.

La herramienta ofrece múltiples funciones, entre las que se destacan:

- Incorporación de nuevos registros
- Publicación de materiales
- Gestión de autoridades
- Gestión de lugares
- Gestión de derechos
- Servicio de referencia
- Generación de informes y reportes
- Generación de metadatos EAD, MARCXML, MODS, Dublin Core, y METS
- Exportaciones

4. Prototipo elegido para probar en los RI gestionados en PREBI-SEDICI

La revisión de los tres modelos precedentes y, en especial, el análisis de los flujos de trabajo y la implementación bajo DSpace de los repositorios en gestión (SEDICI y CIC Digital), determinó la decisión de elegir una estructura que se asemeja a la de la Universidad de Michigan. Las tareas de prueba están en sus albores: al momento se ha instalado en un servidor la aplicación Archivematica y la funcionalidad TRAC (Trustworthy Repositories Audit and Certification). Dado que estos elementos son parte de la experiencia que se quiere compartir, es que se describe con mayor amplitud lo relativo a Archivematica a continuación.

4.1. Instalación

La instalación de Archivematica requiere un servidor GNU/Linux (de momento sólo están soportados Ubuntu Server 14.04.5 y CentOS 7.3.1611, ambos de 64 bits), MySQL (también soporta Percona y MariaDB), un servidor HTTP (Nginx o Apache) y ElasticSearch. La guía de instalación disponible en el sitio web de Archivematica (Archivematica, 2017a) sugiere los requerimientos de hardware mínimos tanto para entornos de prueba como para entornos de producción, y también detalla la secuencia de pasos necesaria para realizar una instalación típica, tanto en Ubuntu Server como en CentOS. La instalación de Archivematica incluye también la instalación del Servicio de Almacenamiento (StorageService), un software web encargado de la gestión de los espacios de almacenamiento disponibles para acceder desde Archivematica. En este software se configuró un directorio local para el proceso de *Transfer* (proceso de Archivematica previo al *Ingest*). Se generaron también usuarios del sistema con permisos de lectura y escritura por medio de SFTP en dicho directorio. Esto permite agilizar la carga de archivos y estructuras de directorios completas, y así realizar las pruebas desde cualquier computadora con un cliente de este protocolo.

4.2. Justificación de la elección, pros y contras

Una característica esencial de Archivematica está vinculada a su diseño, que incluye herramientas ya probadas para realizar las distintas funciones sugeridas en el modelo abstracto OAIS. Esto se consideró, además de una diferencia, una ventaja comparativa en relación a modelos como el SPE, ya que en una arquitectura única se realizan las funciones de preservación esperadas; eso, por otra parte, dificulta el seguimiento de los distintos pasos.

Una cuestión particularmente importante en relación a los objetivos de este trabajo, y que comunica las necesidades de los repositorios gestionados por el grupo PREBI-SEDICI, está vinculada a verificar las capacidades de los repositorios SEDICI y CIC Digital en términos de confiabilidad, atendiendo al Audit and Certification of Trustworthy of Digital Repositories (CCSDS, 2011) que luego se plasma en la ISO 16363. En este sentido, Archivematica resulta también apto al incorporar la herramienta desarrollada por el MIT en el proyecto dedicado a brindar servicios de curaduría y preservación, que se denomina TRAC Review Tool, y que se instala de manera independiente a Archivematica.

4.3. Funcionalidad básica de Archivematica

Archivematica ofrece un conjunto integrado de herramientas libres y de código abierto que permite a los usuarios procesar OD desde su ingesta hasta su almacenamiento, archivo y acceso en conformidad con el modelo funcional ISO-OAIS y otros estándares de conservación digital y buenas prácticas.

La estructura de Archivematica[1] se basa en dos elementos fundamentales y a la vez complementarios: los microservicios y las *Foss tools;* a través de ellos, este desarrollo permite dar cumplimiento del modelo OAIS. Estas herramientas se integran en los diferentes módulos de la plataforma y es posible actualizarlas y configurarlas individualmente. Gracias a ellas se lleva a cabo la normalización de los diferentes formatos de archivo a lo largo del flujo de trabajo, que se inicia con la transferencia de la información al sistema.

Los microservicios son procesos utilizados para llevar a cabo tareas, acciones y transformaciones durante el procesamiento de los paquetes de información en cada uno de los estados del proceso de gestión de los archivos digitales (transferencia, ingesta, depósito y acceso). El administrador tiene la posibilidad de personalizarlos y distribuirlos a lo largo del flujo de trabajo en función de sus necesidades. Algunas de estas acciones realizadas sobre los archivos son automáticas, mientras que otras pueden requerir la intervención del responsable del repositorio que deberá tomar decisiones, en muchos casos estratégicas.

La comunicación entre Archivematica y el administrador en los diversos procesos se realiza a través de un Escritorio (*Dashboard)* que despliega los microservicios y en algunos casos solicita aprobación o atención del administrador. Este Escritorio tiene la apariencia de la Figura 5:

[1] Archivematica ofrece en su sitio una amplia documentación; dada la versión instalada para las pruebas, se ha consultado el manual de la versión 1.5, disponible en https://www.Archivematica.org/es/docs/Archivematica-1.5/

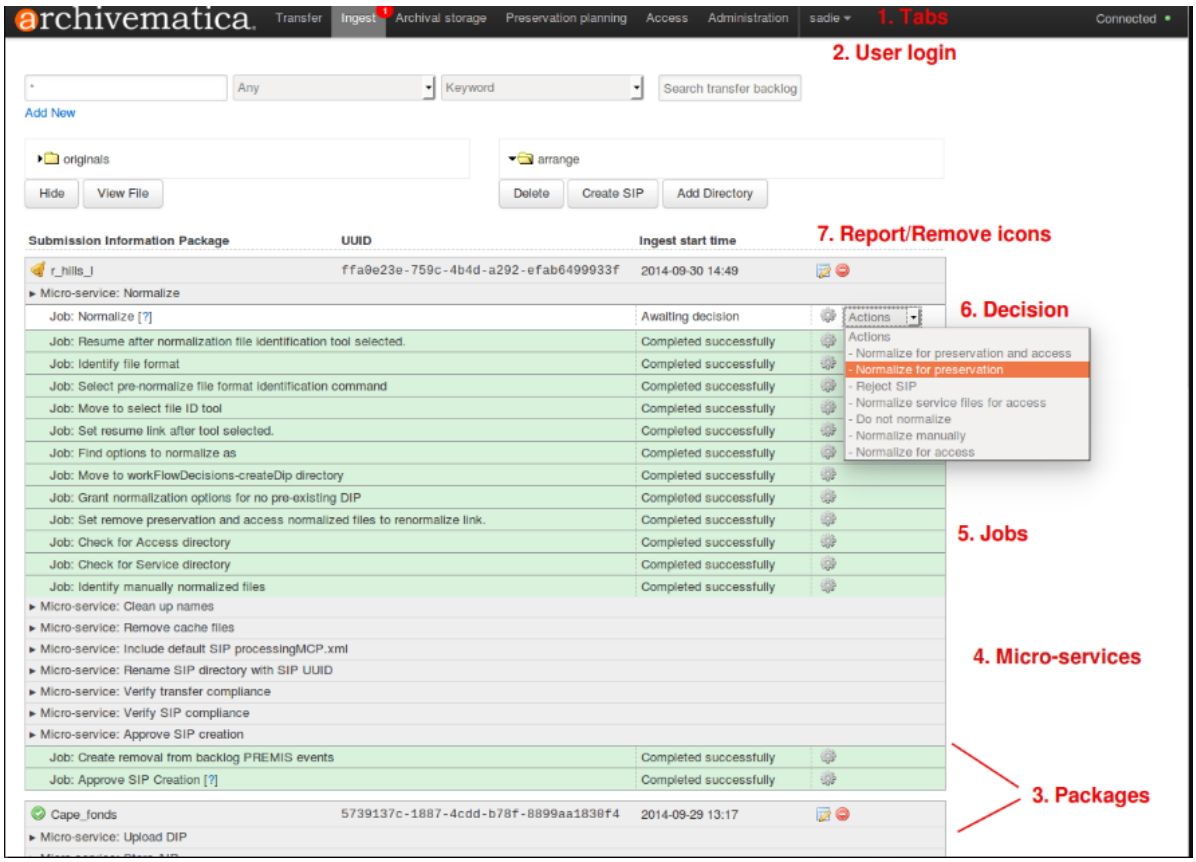

Figura 5. Escritorio de Archivematica

Fuente: https://www.archivematica.org/es/docs/archivematica-1.5/_images/Dashboard.png

Archivematica usa los esquemas de metadatos PREMIS, METS y Dublin Core, pero también permite la importación de otros metadatos que el administrador hubiera agregado al OD. Archivematica implementa planes de preservación para diferentes tipos de contenido. Al momento que se instala, realiza una conexión con los Registros de Políticas de Formato (Format Policy Registry, FPR) para actualizar su base de datos local. Este registro permite a los usuarios de Archivematica definir las políticas para los distintos formatos de archivos.

El Escritorio de Archivematica permite seguir las acciones que suceden en los distintos procesos y microservicios, así como también hacer un seguimiento de los eventos, estados y errores que se producen.

A diferencia del modelo OAIS, en el que el flujo de trabajo comienza en la ingesta, Archivematica pone de manifiesto el proceso previo, llamado Transferencia (*Transfer*), que es un proceso que transforma en un SIP (Submission Information Package, en términos de OAIS), verifica y valida un conjunto de OD (incluso directorios enteros), archivos que vienen de un repositorio DSpace o de otras aplicaciones como puede verse en los dos

modelos propuestos. El administrador elige la opción adecuada y de ese modo se inicia el proceso de transferencia: por ejemplo, puede seleccionar un directorio donde tiene preparados los contenidos a ser sometidos al proceso de curación de Archivematica; estos contenidos pueden incluir documentos de acuerdo con los proveedores de contenidos, en cuyo caso el administrador deberá crear los directorios necesarios para organizarlos. Para realizar las primeras pruebas, se recomienda elegir un conjunto acotado de OD, con algunos formatos en desuso o archivos malformados, a fin de familiarizarse con los reportes entregados por la herramienta. La aprobación del *Transfer* es un proceso manual, una vez que el administrador dispara la acción de validación de la entrega (primer microservicio) en el Escritorio.

Los procesos que se han ejecutado en esta fase de pruebas son:

- Aprobar la entrega
- Comprobar el cumplimiento de requisitos
- Renombrar los archivos añadiendo un identificador único
- Crear sumas de verificación y comprobarlas
- Crear archivos METS XML
- Colocar los archivos en cuarentena si es necesario
- Identificar formatos de los archivos
- Extraer archivos empaquetados
- Comprobar la existencia de virus
- Mover los archivos al directorio de entregas finalizadas
- Crear el SIP[2]

Una gran variedad de herramientas se encargan de los procesos precedentes. Por ejemplo, se utiliza Bagit para el empaquetado y almacenamiento de los objetos digitales, Fido y Siegfried para la extracción de metadatos, Jhove y FITS para la extracción y validación, entre otras. Esto simplifica mucho el trabajo del administrador, porque si bien se puede observar el avance del sistema a través de los distintos pasos que se van ejecutando, no es necesario conocer en detalle cada una de las herramientas.

[2] Como alternativa pueden enviarse los archivos al *backlog* para su posterior procesamiento e incluso crear reportes, como puede verse en la documentación de Archivematica, disponible en https://www.archivematica.org/es/docs/archivematica-1.5/user-manual/transfer/manage-backlog/#retrieve-from-backlog

Llegado a este punto se activa la función de Ingesta (*Ingest*), que se encuentra en el menú de Archivematica. *Ingest* despliega otros microservicios que actuarán sobre el SIP, a saber:

- Normalización
- Agregado de metadatos (esto puede ser previo o posterior a la normalización)
- PREMIS-Derechos

La normalización es el proceso de convertir los OD suministrados a formatos de preservación y/o acceso; es necesario resaltar que los objetos originales siempre se mantienen junto con sus versiones normalizadas. La normalización para preservación y acceso crea copia para tener un objeto preservable (AIP) y un objeto entregable (DIP). Una vez aprobada la normalización, el SIP atraviesa una serie de microservicios, incluyendo el procesamiento de la documentación de la presentación, la generación del archivo METS, la indexación, la generación del DIP y el empaquetado del AIP.

Como en cada paso, es posible revisar los resultados y verificar que todo esté correcto. Una vez cumplida la normalización, el sistema preguntará si se desea guardar el AIP y publicar el DIP, e incluso si se precisa una revisión del AIP (ver Figura 6).

Figura 6. Captura de pantalla del Escritorio de Archivematica y revisión de AIP

Fuente: https://www.archivematica.org/es/docs/archivematica-1.5/user-manual/ingest/ingest/#store-aip

Es posible generar un proceso de revisión del AIP si se considera necesario. El manual recomienda revisar y almacenar el AIP antes de publicar el DIP porque si ocurriera algún problema con el AIP habría que localizar el DIP y borrarlo. Archivematica soporta la subida de DIP a AtoM, ArchivesSpace,

CONTENTdm y Archivist's Toolkit. También permite la re-ingesta de un AIP para el agregado de metadatos.

Las funciones del Archival Storage no requieren demasiados detalles; sin embargo, son válidas algunas aclaraciones. Archivematica utiliza una estructura de árbol de directorios para almacenar los AIP localmente. La estructura en árbol está basada en los identificadores persistentes (16 bits) del AIP; también permite múltiples sitios de almacenamiento locales o remotos e incluso localizaciones LOCKSS. El detalle del procedimiento de integración se encuentra en la wiki de Archivematica (Archivematica 2017b).

La pestaña Archival Storage del Escritorio muestra una tabla con información sobre los AIP almacenados y allí el administrador puede ordenarlos o copiarlos. La identificación de cada AIP se realiza a partir de su nombre y el identificador asignado durante la formación del SIP. La Figura 7 muestra una vista de dos archivos y su identificación:

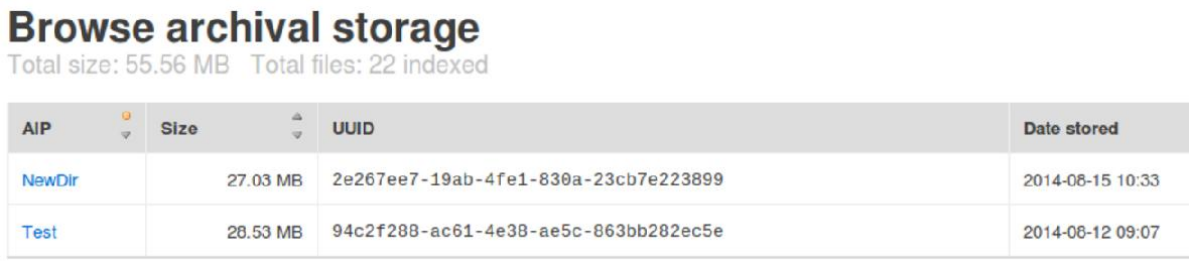

Figura 7. Captura de pantalla del Escritorio con el listado de archivos

Fuente: https://www.archivematica.org/es/docs/archivematica-1.5/_images/ArchStorTab1.png

En relación al módulo de planeamiento de la preservación es oportuno decir que uno de sus principales cometidos (como ya se esbozó en el proceso de ingesta) es la normalización de los archivos para atender tanto a la preservación como al acceso. Al realizar la primera conexión con el FPR, Archivematica puede intercambiar datos sobre el agente y su identificación así como el identificador único del archivo y la dirección de IP, y asimismo el tiempo de realización del evento.

Cuando se crea una nueva versión de formato debe existir un texto que describa el formato a la manera de un archivo METS; el número de versión del formato específico que se esté tratando de curar; el ID de PRONOM, es decir el identificador único de la versión del formato específico en PRONOM, el registro del formato de los Archivos Nacionales del Reino Unido, y también una indicación sobre si el formato es adecuado para acceso

y/o preservación. Si bien Archivematica soporta una amplia gama de formatos, no siempre normaliza todos los formatos, como sucede con MS Word[3].

El módulo de planeamiento de la preservación de Archivematica (Borthwick Institute for Archives, 2017) tiene un elemento central que es la Tabla de Planeamiento de la Preservación donde se despliega el FPR local y donde el administrador puede agregar formatos o editar los existentes al realizar la primera conexión. La Figura 8 muestra la identificación del formato PDF/A.

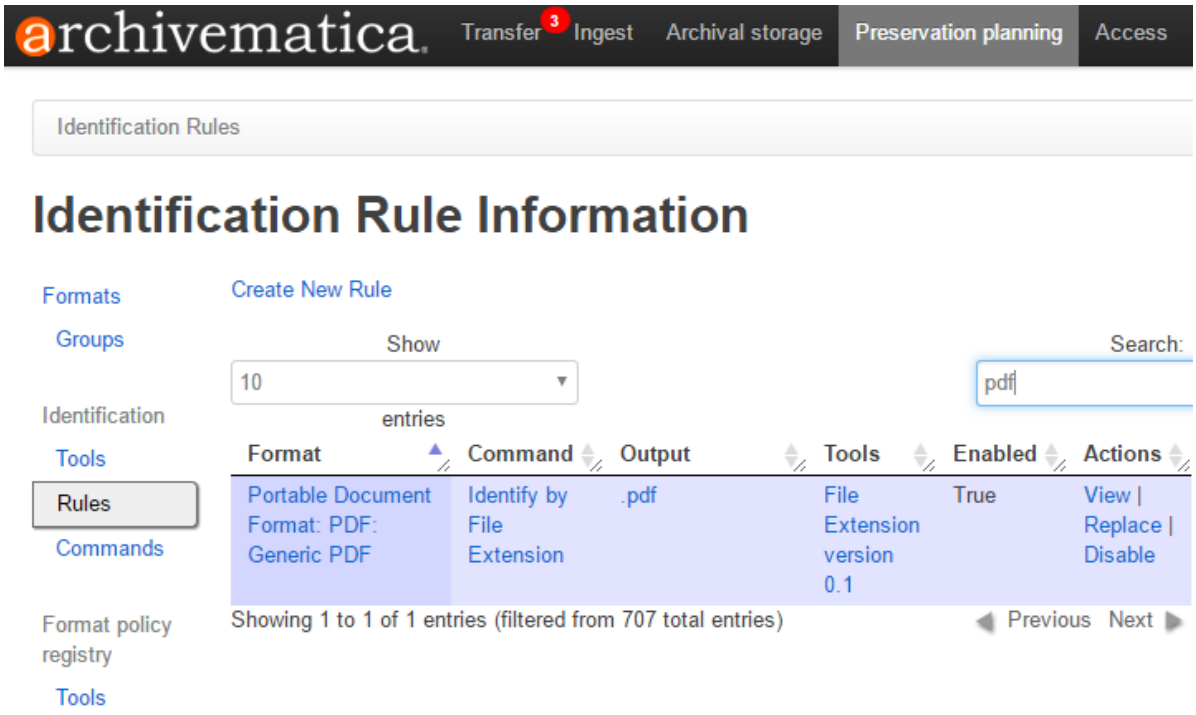

Figura 8. Captura de pantalla del Escritorio de Preservation Planning

Las reglas vigentes del FPR pueden actualizarse en cualquier momento por parte del administrador. Dentro de este módulo también es posible ver lo relativo a identificación de formatos y contraste con el FPR; para este proceso se utilizan distintas herramientas como FIDO (Open Planets Foundation) que contrasta con el registro PRONOM; un *script* que identifica por extensión del archivo y Siegfried que también trabaja con PRONOM. La

[3] Como se especifica en la documentación de Archivematica (2017c), "algunos formatos, como documentos de Microsoft Word, no tienen el mejor formato de preservación pero de todos modos son localizables y no necesitan ser normalizados en la actualidad. En estos casos, el procedimiento estándar de Archivematica es dejarlos en su formato original" (traducción propia).

versión 1.5 cuenta con cinco herramientas para la caracterización de los formatos, entre ellas, FITS. Para el proceso de validación se utiliza Jhove.

TRAC Review Tool

La herramienta TRAC Review[4] es un desarrollo del MIT, basado en el sistema de gestor de contenidos Drupal, que es de gran ayuda para las organizaciones que desean implementar un repositorio confiable, y resulta en particular útil para aquellas instituciones que utilizan Archivematica para esta tarea. TRAC permite realizar el seguimiento y dejar constancia acerca del cumplimiento (o falta de cumplimiento) de los requerimientos listados por el CCSDS mencionado previamente, que luego fuera aprobado como norma ISO 16363 (2012) y que se basa, precisamente, en los requerimientos provistos por TRAC. Esta autoevaluación permite demostrar las buenas prácticas y la conformidad para con la comunidad a la que está dedicado un repositorio. TRAC propone un gran cantidad de responsabilidades, que en muchas organizaciones se distribuyen entre diferentes comisiones o unidades responsables de determinados requerimientos.

Trabajos y tareas pendientes

Al plantear un sistema de preservación digital, es importante conocer la experiencia de otras organizaciones y las combinaciones de herramientas que han resultado exitosas, no sólo porque cumplen con los requerimientos técnicos para las que fueron concebidas, sino también porque permiten combinarse con otras herramientas a fin de implementar sistemas complejos que se adecúan a las necesidades de las organizaciones que los utilizan. La tarea realizada hasta el momento en PREBI-SEDICI ha sido meramente la de entrar en conocimiento de las herramientas que integran la estructura de preservación seleccionada y describir tres casos de éxito reconocidos internacionalmente: el proyecto SCAPE, la arquitectura propuesta por la Universidad British Columbia y la estructura utilizada por la Bentley Historical Library de la Universidad Michigan. En muchos aspectos aún falta tener en claro la complejidad que representa Archivematica, utilizado en dos de las tres propuestas analizadas, y, adicionalmente, analizar la generación de

[4] Pueden verse sus requerimientos de instalación en https://www.archivematica.org/en/docs/archivematica-1.5/user-manual/getting-started/trac/

las conexiones con ArchivesSpace y un repositorio de pruebas implementado en DSpace sobre el que será necesario habilitar una conexión bajo el protocolo SWORD 2. Hasta aquí sólo se tienen pruebas aisladas sobre Archivematica y un gran trabajo pendiente. Sin embargo, el análisis de estas herramientas se considera un gran avance para el futuro de los repositorios en gestión y la validez de estas notas tiene como justificación, como se enunciara, la de compartir estos primeros pasos buscando la mejor solución al creciente problema de la preservación de archivos digitales.

Bibliografía

ARCHIVEMATICA (2017a). Installation | Documentación (Archivematica 1.6) | Archivematica: open-source digital preservation system. Disponible en <https://www.archivematica.org/es/docs/archivematica-1.6/admin-manual/installation/installation/>

ARCHIVEMATICA (2017b). [sitio web] <https://wiki.Archivematica.org/LOCKSS_Integration>

ARCHIVEMATICA (2017c). Preservation planning. Disponible en <https://www.archivematica.org/fr/docs/archivematica-1.4/user-manual/preservation/preservation-planning/>

ASHLEY, LORI J. (2016). "Theory: Creating a preservation strategy". En Bantin, Philip C. (ed.). *Building Trustworthy Digital Repositories: Theory and Implementation*. Rowman & Littlefield: London.

BORTHWICK INSTITUTE FOR ARCHIVES (2017). Filling the Digital Preservation Gap. Report on Archivematica for research data now available. Disponible en <https://www.york.ac.uk/borthwick/projects/archivematica/>

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS (CCDS) (2011). *Audit and certification of Trustworthy Digital Repositories*. The Magenta Book. Disponible en <https://public.ccsds.org/pubs/652x0m1.pdf>

DE GIUSTI, MARISA R.; LIRA, ARIEL; VILLARREAL, GONZALO; TEXIER, JOSÉ (2012). "Las actividades y el planeamiento de la preservación en un repositorio institucional". En BIREDIAL 2012, Barranquilla (Colombia). Disponible en <http://sedici.unlp.edu.ar/handle/10915/26045>

DE GIUSTI, MARISA R. (2016). "Las dificultades de la preservación digital: problemas, desafíos y propuestas para los repositorios". En BIREDIAL-ISTEC 2016, San Luis Potosí (México). Disponible en <http://sedici.unlp.edu.ar/handle/10915/56288>

ECKARD, MAX; PILLEN, DALLAS AND SCHALLCROSS, MIKE (2017). "Bridging Technologies to Efficiently Arrange and Describe Digital Archives: the Bentley

Historical Library's ArchivesSpace-Archivematica-DSpace Workflow Integration Project". *Code(4)Lib Journal*, 35, 1-30. Disponible en <http://journal.code4lib.org/articles/12105>

FARIA, LUÍS; FERREIRA, MIGUEL; CASTRO, RUI; BARBEDO, FRANCISCO; HENRIQUES, CECÍLIA; CORUJO, LUÍS; RAMALHO, JOSÉ CARLOS (2009). "RODA: a service-oriented repository to preserve authentic digital objects". En International Conference on Open Repositories (OR 2009), 4, Atlanta, USA. Disponible en <http://repositorium.sdum.uminho.pt/handle/1822/9408>

FEDORA (2017). [sitio web] Fedora Commons, University of Virginia and Cornell University. <http://www.fedora.info/>

FERREIRA, MIGUEL; FARIA, LUÍS; HAHN, MATTHIAS; DURETEC, KRESIMIR (2014). "SCAPE: Report on compliance validation". Whitepaper. Disponible en <http://hdl.handle.net/1822/30689>

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO) (2012). ISO 14721. Space data and information transfer systems - Open archival information system (OAIS) - Reference model. <https://www.iso.org/standard/57284.html>

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO) (2012). ISO 16363. Space data and information transfer systems - Audit and certification of trustworthy digital repositories. Disponible en <https://www.iso.org/standard/56510.html>

PITTI, DANIEL V. (1999). "Encoded Archival Description. An Introduction and Overview". *D-Lib Magazine*, 5(11), November. DOI: 10.1045/november99-pitti

SPROUT, BRONWEN AND ROMKEY, SARAH (2016). "Implementation: Building a preservation strategy around Archivematica". En Bantin, Philip C. (ed.). *Building Trustworthy Digital Repositories: Theory and Implementation*. Rowman & Littlefield: London.